Appendix A. Proof of Lemma 3

First, we show that the Markov chain $\{X_n^m(t), t \geq 0\}$ is ergodic. It is irreducible and aperiodic since any state can reach the initial state $X_n^m(0) = 0$ via update and the initial state can also reach to itself. It is positive recurrent since the expected return time to state 0 is finite under the condition in Eq. (5). By the ergodicity of Markov chain, we have that the limiting distribution $\pi = (\pi_0, \pi_1, \ldots)$ exists and it is also the unique stationary distribution. Here,

$$\pi_k = \lim_{n\to\infty} \mathbb{P}_{vk}^n \qquad \forall v$$

where $\mathbb{P}_{vk}^n$ is the probability of being in state k in n steps, given we are in state v now. Since the limiting distribution is independent of the initial state, we can pick any state to start with. In particular, we pick the initial state as $X_n^m(0) = 0$. By Eq. (9), we have for any t

$$\mathbb{E}\left[X_n^m(t+1)\right] \leq (1-p)\mathbb{E}\left[X_n^m(t)\right] + \lambda_\Sigma + \mu_{max}. \tag{A.1}$$

We conduct an inductive process to establish that $\mathbb{E}\left[X_n^m(t)\right] \leq \frac{\lambda_\Sigma + \mu_{max}}{p}$ for all t. First, the basis is true since $X_n^m(0) = 0$. We assume $\mathbb{E}\left[X_n^m(t_0)\right] \leq \frac{\lambda_\Sigma + \mu_{max}}{p}$ holds for $t = t_0$, then by Eq. (A.1), we have $\mathbb{E}\left[X_n^m(t_0+1)\right] \leq \frac{\lambda_\Sigma + \mu_{max}}{p}$ also holds and hence completing the proof. In particular, $\mathbb{E}\left[X_n^m(t)\right] \leq \frac{\lambda_\Sigma + \mu_{max}}{p} \leq \frac{2\mu_\Sigma}{p}$ for all t. This directly means that the stationary distribution π of $\{X_n^m(t), t \geq 0\}$ has a finite bound on the mean (independent of ϵ). Similarly, by applying the same step as in Eq. (9) combined with the inductive argument, we can obtain that all the moments of the stationary distribution are bounded, because all the moments of the total arrival and each service are bounded (independent of ϵ) by our light-tailed assumption.

Let $f_T \triangleq \frac{1}{T}\left(\sum_{t=t_0}^{t_0+T-1} X_n^m(t) \mid Z(t_0) = Z\right)$. By ergodicity (i.e., time-average = ensemble-average), we have that for any starting point $Z(t_0) = Z$, with probability 1 such that

$$\lim_{T\to\infty} f_T = \sum_{k=0}^{\infty} k\pi_k \leq \frac{2\mu_\Sigma}{p}.$$

As a result, we can find a finite T_1 (independent of ϵ since all the moments of π are bounded with independence of ϵ) such that for all $T \geq T_1$, $f_T \leq \frac{4\mu_\Sigma}{p} \triangleq L$ with probability 1. Therefore, if $T \geq T_1$,

$$\mathbb{E}\left[\sum_{t=t_0}^{t_0+T-1} X_n^m(t) \mid Z(t_0) = Z\right] \leq LT,$$

which completes the proof.

Appendix B. Proof of Lemma 4

Consider the left-hand-side (LHS) of the inequality in Lemma 4.

$$
\begin{aligned}
LHS &\overset{(a)}{\leq} \sum_{t=t_0}^{t_0+T-1} \sum_{m=1}^{M} \lambda_m \delta \mathbb{E}\left[\widetilde{Q}^m_{min}(t) - \widetilde{Q}^m_{max}(t) \mid Z(t_0) = Z\right] \\
&\overset{(b)}{\leq} \sum_{m=1}^{M} \lambda_m \delta \mathbb{E}\left[\widetilde{Q}^m_{min}(t_0+1) - \widetilde{Q}^m_{max}(t_0+1) \mid Z(t_0) = Z\right] \\
&\quad + \sum_{m=1}^{M} \lambda_m \delta \mathbb{E}\left[\widetilde{Q}^m_{min}(t_0+2) - \widetilde{Q}^m_{max}(t_0+2) \mid Z(t_0) = Z\right],
\end{aligned}
$$

where (a) follows from the definition of δ-tilted sum condition and a decomposition similar to Eqs. (17) and (18). Specifically, note that $\sum_{n=1}^{N} \bar{Q}^m(t)\beta_n^m(t) = 0$ for each m. (b) holds since all the terms in the summation are non-positive and $T \geq 3$.

Now, we define random variable $\mathcal{I}^m_{min}(t)$ as the indicator function which obtains 1 if the server with the minimal true queue length at the end of time-slot t is updated by dispatcher m. Similarly, $\mathcal{I}^m_{max}(t)$ is defined as an indicator function which obtains 1 if the server with the maximal true queue length at the end of time-slot t is updated by dispatcher m. In the following, we consider the event that $\mathcal{I}^m_{max}(t_0) = 1$ and $\mathcal{I}^m_{min}(t_0+1) = 1$, i.e., at the end of time-slot t_0 the server with the maximal actual queue is updated and at the end of time-slot t_0+1 the server with the minimal actual queue is updated. Then, by law of expectation, we have

$$
\begin{aligned}
LHS \leq &\sum_{m=1}^{M} \lambda_m \delta p^2 \mathbb{E}\Big[\widetilde{Q}^m_{min}(t_0+1) - \widetilde{Q}^m_{max}(t_0+1) \\
&\widetilde{Q}^m_{min}(t_0+2) - \widetilde{Q}^m_{max}(t_0+2) \mid Z, \mathcal{I}^m_{max}(t_0) = 1, \mathcal{I}^m_{min}(t_0+1) = 1\Big] \\
\overset{(a)}{\leq} &\sum_{m=1}^{M} \lambda_m \delta p^2 \mathbb{E}\Big[\widetilde{Q}^m_{min}(t_0+1) - Q_{max}(t_0+1) \\
&Q_{min}(t_0+2) - \widetilde{Q}^m_{max}(t_0+2) \mid Z, \mathcal{I}^m_{max}(t_0) = 1, \mathcal{I}^m_{min}(t_0+1) = 1\Big] \\
\overset{(b)}{\leq} &\sum_{m=1}^{M} \lambda_m \delta p^2 \mathbb{E}\left[-Q_{max}(t_0+1) + Q_{min}(t_0+2)\right. \\
&\left.\mid Z(t_0) = Z, \mathcal{I}^m_{max}(t_0) = 1, \mathcal{I}^m_{min}(t_0+1) = 1\right],
\end{aligned}
\tag{B.1}
$$

where (a) follows from the fact that with $\mathcal{I}^m_{max}(t_0) = 1$, $\widetilde{Q}^m_{max}(t_0+1) \geq Q_{max}(t_0+1)$ and with $\mathcal{I}^m_{min}(t_0+1) = 1$, $\widetilde{Q}^m_{min}(t_0+2) \leq Q_{min}(t_0+2)$; (b) holds since $\widetilde{Q}^m_{min}(t_0+1) \leq \widetilde{Q}^m_{max}(t_0+1) \leq \widetilde{Q}^m_{max}(t_0+2)$. This is because under any LED policy the decrease of local estimate can only happen when the queue is updated.

In order to relate the queue lengths at time-slots $t_0 + 1$ and $t_0 + 2$ to the queue length at t_0, we use the following result.

Claim 1. *For any t_0, we have*

1. $\mathbb{E}\left[Q_{min}(t_0+1) \mid Z(t_0) = Z\right] \leq Q_{min}(t_0) + M_1$
2. $\mathbb{E}\left[Q_{max}(t_0+1) \mid Z(t_0) = Z\right] \geq Q_{max}(t_0) - M_1$

where $M_1 = \mu_\Sigma$.

Proof. Let us start with the first result. Suppose that the server i is the server with the shortest queue length at time-slot t_0. We have

$$\begin{aligned}
\mathbb{E}\left[Q_i(t_0+1) \mid Z(t_0) = Z\right] &= \mathbb{E}\left[Q_i(t_0) + A_i(t_0) - S_i(t_0) + U_i(t) \mid Z(t_0)\right] \\
&\leq Q_i(t_0) + \max(\lambda_\Sigma, \mu_i) \\
&\leq Q_i(t_0) + \mu_\Sigma \\
&= Q_i(t_0) + M_1. \qquad \text{(B.2)}
\end{aligned}$$

If at time-slot $t_0 + 1$, the same server i is still the one with the shortest queue length, then we are done. If not, suppose that some other server j is the one with the shortest queue length at time-slot $t_0 + 1$. Now, we assume that $\mathbb{E}\left[Q_j(t_0+1) \mid Z(t_0) = Z\right] > Q_i(t_0) + M_1$, and we hope to arrive at a contradiction.

First, since $Q_j(t_0+1) \leq Q_i(t_0+1)$, $\mathbb{E}\left[Q_j(t_0+1) \mid Z(t_0) = Z\right] \leq \mathbb{E}\left[Q_i(t_0+1) \mid Z(t_0) = Z\right]$. Combined with our assumption, we get

$$Q_i(t_0) + M < \mathbb{E}\left[Q_j(t_0+1) \mid Z(t_0) = Z\right] \leq \mathbb{E}\left[Q_i(t_0+1) \mid Z(t_0) = Z\right] \qquad \text{(B.3)}$$

Thus, we can see Eq. (B.3) contradicts with Eq. (B.2). Hence,

$$\mathbb{E}\left[Q_j(t_0+1) \mid Z(t_0) = Z\right] \leq Q_i(t_0) + M_1,$$

which finishes the proof of the first result. Same argument can be applied to prove the second result. □

Now, we are ready to bound Eq. (B.1). First,

$$LHS \overset{(a)}{\leq} \sum_{m=1}^{M} \lambda_m \delta p^2 \mathbb{E}\left[-Q_{max}(t_0+1) \mid Z(t_0) = Z\right] + \qquad \text{(B.4)}$$

$$\sum_{m=1}^{M} \lambda_m \delta p^2 \mathbb{E}\left[Q_{min}(t_0+2) \mid Z(t_0) = Z, \phi(t_0)\right], \qquad \text{(B.5)}$$

where $\phi(t_0) \triangleq \mathcal{I}^m_{max}(t_0) = 1, \mathcal{I}^m_{min}(t_0+1) = 1$. The inequality follows from the fact that given $Z(t_0)$, $Q_{max}(t_0+1)$ is independent of $\mathcal{I}^m_{max}(t_0)$ and $\mathcal{I}^m_{min}(t_0+1)$.

By using the bound in Claim 1, we have an upper bound for the term in Eq. (B.4)

$$\sum_{m=1}^{M} \lambda_m \delta p^2 \mathbb{E}\left[-Q_{max}(t_0+1) \mid Z(t_0)=Z\right]$$
$$\leq \sum_{m=1}^{M} \lambda_m \delta p^2 (-Q_{max}(t_0) + M_1). \tag{B.6}$$

For the term in Eq. (B.5), we have

$$\begin{aligned}
&\sum_{m=1}^{M} \lambda_m \delta p^2 \mathbb{E}\left[Q_{min}(t_0+2) \mid Z(t_0)=Z, \phi(t_0)\right] \\
\stackrel{(a)}{=} &\sum_{m=1}^{M} \lambda_m \delta p^2 \mathbb{E}\left[\mathbb{E}\left[Q_{min}(t_0+2) \mid Z(t_0+1)\right] \mid Z(t_0)=Z,\right. \\
&\left.\mathcal{I}_{max}^m(t_0)=1, \mathcal{I}_{min}^m(t_0+1)=1\right]
\end{aligned}$$

where (a) follows from the tower property of conditional expectation and the fact that given $Z(t_0+1)$, $Q_{min}(t_0+2)$ is independent of $\phi(t_0)$. Then, it can be upper bounded as follows.

$$\begin{aligned}
&\sum_{m=1}^{M} \lambda_m \delta p^2 \mathbb{E}\left[Q_{min}(t_0+2) \mid Z(t_0)=Z, \phi(t_0)\right] \\
\stackrel{(a)}{\leq} &\sum_{m=1}^{M} \lambda_m \delta p^2 \mathbb{E}\left[Q_{min}(t_0+1) + M_1 \mid Z(t_0)=Z,\right. \\
&\left.\mathcal{I}_{max}^m(t_0)=1, \mathcal{I}_{min}^m(t_0+1)=1\right] \\
\stackrel{(b)}{=} &\sum_{m=1}^{M} \lambda_m \delta p^2 \mathbb{E}\left[Q_{min}(t_0+1) + M_1 \mid Z(t_0)=Z\right] \\
\stackrel{(c)}{\leq} &\sum_{m=1}^{M} \lambda_m \delta p^2 (Q_{min}(t_0) + 2M_1),
\end{aligned}$$

where (a) comes from the bound in Claim 1; (b) holds since given $Z(t_0)$, $Q_{min}(t_0+1)$ is independent of the event $\mathcal{I}_{max}^m(t_0)=1, \mathcal{I}_{min}^m(t_0+1)=1$; (c) holds by the bound in Claim 1 again.

Thus, combining the bounds for Eqs. (B.4) and (B.5), yields

$$\begin{aligned}
LHS &\leq \sum_{m=1}^{M} \lambda_m \delta p^2 (Q_{min}(t_0) - Q_{max}(t_0) + 3M_1) \\
&= \lambda_\Sigma \delta p^2 (Q_{min}(t_0) - Q_{max}(t_0) + 3M_1 \lambda_\Sigma \delta p^2 \\
&\leq -\frac{\lambda_\Sigma \delta p^2}{\sqrt{N}} \left\|\mathbf{Q}_\perp(t_0)\right\| + 3(\mu_\Sigma)^2 p^2,
\end{aligned} \tag{B.7}$$

in which the last inequality follows from the fact that $\|\mathbf{Q}_\perp(t_0)\| \leq \sqrt{N}\,(Q_{max}(t_0) - Q_{min}(t_0))$ and $M_1 = \mu_\Sigma$ with $\delta \leq 1$. Hence, the proof of Lemma 4 is complete.

Appendix C. Proof of Lemma 5

First, note that by Eq. (22), we have

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{Q}_\perp(t+1)\| \mid Z(t) = Z\right] \\
&\leq \|\mathbf{Q}_\perp(t)\| + 2\mathbb{E}\left[\|\mathbf{A}(t)\| + 2\,\|\mathbf{S}(t)\| \mid Z\right] \\
&\leq \|\mathbf{Q}_\perp(t)\| + 2\mathbb{E}\left[\sqrt{NA_\Sigma^2(t)} + 2\sqrt{NS_\Sigma^2(t)} \mid Z\right] \\
&\overset{(a)}{\leq} \|\mathbf{Q}_\perp(t)\| + 2\sqrt{N(\sigma_\Sigma^2 + \mu_\Sigma^2)} + 4\sqrt{N(\nu_\Sigma^2 + \mu_\Sigma^2)} \\
&\overset{(b)}{\leq} \|\mathbf{Q}_\perp(t)\| + C_2,
\end{aligned} \tag{C.1}$$

where (a) follows from Jensen's inequality for concave function; in (b) C_2 is a finite constant independent of ϵ, which holds by our light-tailed assumption.

Now, by using the result above, we have

$$\begin{aligned}
&\sum_{t=t_0}^{t_0+T-1} \mathbb{E}\left[\sum_{n=1}^{N} Q_{\perp,n}(t)\frac{-\epsilon\mu_n}{\mu_\Sigma} \mid Z(t_0) = Z\right] \\
&\overset{(a)}{\leq} \sum_{t=t_0}^{t_0+T-1} \mathbb{E}\left[\epsilon\sqrt{N}\,\|\mathbf{Q}_\perp(t)\| \mid Z(t_0) = Z\right] \\
&\leq \epsilon\sqrt{N}\left(\sum_{t=t_0}^{t_0+T-1} \|\mathbf{Q}_\perp(t_0)\| + (t - t_0)C_2\right) \\
&\leq T\epsilon\sqrt{N}\,\|\mathbf{Q}_\perp(t_0)\| + T^2 C_2,
\end{aligned}$$

where (a) is true since $\|\mathbf{x}\|_1 \leq \sqrt{N}\,\|\mathbf{x}\|$ for any $\mathbf{x} \in \mathbb{R}^N$. Hence, the proof is complete.

Appendix D. Proof of Proposition 1

Recall that $\sigma_t(\cdot)$ is a permutation of $(1, 2, \ldots, N)$ which satisfies

$$\widetilde{Q}^m_{\sigma_t(1)}(t) \leq \widetilde{Q}^m_{\sigma_t(2)}(t) \leq \ldots \leq \widetilde{Q}^m_{\sigma_t(N)}(t).$$

Under L-Pod, given a dispatcher m, we have

$$P^m_{\sigma_t(n)} = \frac{\binom{N-n}{d-1}}{\binom{N}{d}}, \quad 1 \leq n \leq N - d + 1$$

and $P^m_{\sigma_t(n)} = 0$ for all $n > N - d + 1$. Then, by the definition of dispatching preference, we have (for ease of exposition, let $\binom{n}{k} = 0$ if $n < k$)

$$
\begin{aligned}
\sum_{n=1}^{j} \Delta_n^m(t) &= \frac{1}{\binom{N}{d}} \sum_{n=1}^{j} \binom{N-n}{d-1} - \frac{1}{\mu_\Sigma} \sum_{n=1}^{j} \mu_{\sigma_t(n)} \\
&= \frac{1}{\binom{N}{d}} \left(\binom{N}{d} - \sum_{n=j+1}^{N} \binom{N-n}{d-1} \right) - \frac{1}{\mu_\Sigma} \sum_{n=1}^{j} \mu_{\sigma_t(n)} \\
&= 1 - \frac{1}{\binom{N}{d}} \binom{N-j}{d} - \frac{1}{\mu_\Sigma} \sum_{n=1}^{j} \mu_{\sigma_t(n)} \\
&\geq 1 - \frac{1}{\binom{N}{d}} \binom{N-j}{d} - \frac{1}{\mu_\Sigma} \sum_{n=1}^{j} \mu_{[n]} \\
&\geq \delta
\end{aligned}
$$